MANCHESTER 1824
The University of Manchester

The University of Manchester Research

Voltage Scaling for 3-D ICs: When, How, and How Much?

DOI:
10.1016/j.mejo.2017.09.005

Document Version
Accepted author manuscript

Link to publication record in Manchester Research Explorer

Citation for published version (APA):
Kalargaris, C., & Pavlidis, V. (2017). Voltage Scaling for 3-D ICs: When, How, and How Much? *Microelectronics Journal*, *69*, 35-44. https://doi.org/10.1016/j.mejo.2017.09.005

Published in:
Microelectronics Journal

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights
Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy
If you believe that this document breaches copyright please refer to the University of Manchester's Takedown Procedures [http://man.ac.uk/04Y6Bo] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.

OPEN ACCESS

Download date:12. Nov. 2022

Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 00 (2017) 1–11

Procedia Engineering

Voltage Scaling for 3-D ICs: When, How, and How Much?

Harry Kalargaris and Vasilis F. Pavlidis

Advanced Processor Technologies Group
School of Computer Science, University of Manchester

Abstract

Three-dimensional (3-D) integration has great potential to improve the power and performance of integrated circuits. Speed improvements in 3-D ICs originate from the reduction of interconnect (*RC*) delay at the critical paths. The decrease in the power of a 3-D circuit is traditionally considered to result from the reduction of the interconnect capacitance and the number of repeaters due to the shorter wirelength. Since the power and performance of a circuit can significantly differ between the two (planar circuits) and three dimensions (multi-tier circuits) at the same operating voltage, this situation provides a voltage headroom to further improve the speed or the power of the 3-D stack. Voltage scaling can be utilized to improve either the performance or the power of 3-D circuits. In this work, emphasis is placed on reducing the operating voltage, as power savings from the reduction of the wirelength are limited due to the non-negligible parasitic capacitance of the through silicon vias (TSVs) that vertically interconnect the tiers. The operating voltage is decreased by exploiting the additional slack that results from the shorter length of the critical nets, such that the performance of the circuit does not change between the two and three dimensions. Guidelines and a timing model based on the logical effort for identifying whether a circuit can benefit from this approach are offered. In addition, a methodology for applying and evaluating voltage reduction in 3-D ICs at the system level is presented. The typical approach of reducing power due to the decreased wire capacitance in 3-D ICs leads to moderate power savings for a variety of circuits (6.6% on average). Alternatively, the proposed approach results in a decrease in power of 22.3% on average. In addition, a decrease of 34.7% in the peak power is observed for a specific case study as compared to the 3-D circuits where voltage supply is not scaled.

© 2011 Published by Elsevier Ltd.

Keywords: 3-D integration, voltage scaling, power, logical effort

1. Introduction

Power and performance bottlenecks due to the long interconnects have become a pressing issue in modern circuits [1]. Three-dimensional (3-D) integration is a promising technology where tiers are stacked and interconnected vertically, supporting high integration densities. The physical proximity of the devices underpinned by the vertical integration has been shown to improve power and performance in 3-D circuits as compared to two-dimensional (2-D) circuits [1]-[5].

Early case studies of 3-D circuits have shown, for example, that partitioning the logic blocks of a core to several tiers improves the performance of the individual blocks or of the overall system. Folding logic blocks, such as the instruction scheduler, in two tiers results in lowering the delay [2]. Black *et al.*, published a 3-D architecture of Intel™ Pentium® 4 in two tiers [3], where the pipelined *RC* delay is reduced. Moreover, improved speed is reported in other circuits, such as memories, due to the reduced wirelength [3]-[5].

Similarly, the decreased interconnect capacitance and the elimination of buffers are typically considered the primary means for saving power in 3-D ICs. However, both the dynamic and static power linearly depend upon the wire capacitance and number of buffer stages. This dependence leads to moderate power improvements in 3-D ICs [3]-[5]. In addition, through silicon vias (TSVs) that vertically interconnect the tiers exhibit non-negligible parasitic capacitance, which can result in buffer insertion before the TSV [6]. Consequently, the decrease in power offered by vertical integration can be severely constrained if only wirelength reduction is considered. Based on this observation, this work follows a different yet efficient way to decrease power by combining the innate traits of 3-D integration with standard low power methods for integrated systems.

A broad gamut of physical design techniques exist to decrease the power in circuits, such as gate level power optimization [8], clock [9] and power gating [10], multi-threshold logic [11], and voltage supply scaling [12]. One of the most efficient techniques for reducing power consumption is the scaling of the

power supply voltage due to the quadratic dependency of the dynamic power on this voltage. Voltage reduction, however, is associated with a penalty in performance. Consequently, supply voltage scaling is considered in tandem with performance [13], where the operating voltage is chosen to satisfy a specific performance level and *vice versa*.

By considering the well-known power-speed interplay, a methodology to reduce power by decreasing the supply voltage is developed. The incurred increase in delay is compensated by exploiting the additional slack produced by the third dimension. This approach is based on several results which demonstrate that the traditional notion of decreasing power in 3-D ICs due to the reduction of interconnect capacitance is not adequate, in particular, if TSVs exhibit non-negligible parasitic capacitance [6]. In other words, the noticeable decrease in *RC* delay, as the resistance of TSV is considerably smaller than in horizontal wires, is exploited to counterweigh the increase in the delay of logic gates that results from reducing the supply voltage. In this way, power is saved without degrading the performance.

However, depending on the characteristics of the paths within a circuit, the power savings from applying this technique to 3-D ICs can greatly vary. Consequently, the critical paths of a design should be carefully considered to evaluate where voltage reduction does not degrade the target performance of the system. Previous works do not offer a systematic analysis, emphasizing the (critical) paths with the longest inter-tier nets [14], [15]. Although intuitive, this practice may not lead to substantial savings in power since, as shown in this work, long wires are only one important constituent for the application of voltage scaling in 3-D ICs.

Starting from a 2-D circuit, a full chip evaluation is utilized to assess whether voltage scaling is applicable, rather than focusing only on the limited set of paths that span more than one tier. As demonstrated in this work, these paths can exhibit a misleading behavior, in particular, if the speed of the circuit is dominated by the gate delay. In addition, the proposed technique enhances power reduction in 3-D ICs while considering the characteristics of TSV-based vertical integration. A timing model considering both the interconnect length and voltage sensitivity of paths and guidelines for identifying when two-dimensional circuits can benefit from this approach are offered. Moreover, a methodology for applying and evaluating voltage reduction in 3-D ICs at the system level is presented. To quantify the potential power gains from this method, several benchmark circuits are investigated.

The remainder of this paper is organized as follows. In Section 2, the main principles of voltage scaling in modern circuits alongside related works to three-dimensional circuits are discussed. The utilization of voltage scaling in 3-D ICs is discussed in Section 3. In Section 4, an enhanced timing model for paths and guidelines for identifying whether a circuit can benefit from this approach are offered. A new methodology integrated in an advanced EDA flow for globally applying voltage scaling to 3-D circuits is presented in Section 5. Related results for several benchmark circuits are presented in Section 6. Some conclusions are drawn in Section 7.

2. Voltage Scaling in Modern Circuits

In this section, the main principles and different forms of voltage scaling in modern circuits are presented. In addition, works related to low power methods for three-dimensional circuits are discussed.

2.1. Main Principles of Voltage Scaling

A traditional objective for integrated systems is the reduction of power consumption without hindering the performance of the system. Voltage scaling is one of the most efficient techniques to achieve this objective as both the power and speed are considered at the same time. The dynamic and static components of the power are, respectively,

$$P_{dyn} = aCV_{DD}^2 f, \tag{1}$$

and

$$P_{static} = V_{DD} I_{leak}, \tag{2}$$

where a is the fraction of gates switching, C is the total switched capacitance, V_{DD} is the supply voltage, f is the operating frequency of the circuit, and I_{leak} is the off-state leakage current. Therefore, voltage supply reduction can lead to significant power gains due to the quadratic relation of voltage to the dynamic power and the linear dependency upon the leakage power.

However, the supply voltage relates to the delay of a circuit as

$$Delay \propto V_{DD} / (V_{DD} - V_{th})^{\beta} \tag{3}$$

where β is an experimentally derived constant specific to the manufacturing technology ($1 < \beta < 2$) and V_{th} is the threshold voltage [16]. A tradeoff therefore exists between power and performance as a reduction in voltage supply increases proportionally the latency of the circuit. Consequently, voltage scaling is always considered along with performance. This situation leads to different strategies of voltage scaling: i) reducing the operating voltage when the power budget is constrained and ii) increasing the operating voltage for a given frequency target, when speed is the primary objective.

2.2. Types of Voltage Scaling

The technique of voltage scaling can be applied to different circuit granularities. At the system level, voltage changes globally across an entire circuit. Alternatively, voltage can be scaled at the block level, individually adapting the voltage of functional circuits blocks (i.e. chip multiprocessors - CMPs) composing a system [16]. In both cases, the same principles (see Section 2.1) of voltage scaling are applicable in terms of voltage and speed.

Another categorization of voltage scaling schemes is based on the mechanism for altering the voltage both at the system and block level. In multi-voltage scaling (MVS) schemes, few and fixed voltage levels are supported and the power management unit (PMU) is responsible to manage these levels [12]. Similarly, dynamic voltage and frequency scaling (DVFS) is an extension of the MVS scheme where a larger number of voltage levels are supported by the PMU and are dynamically switched

to follow the dynamic behavior of the workload. In an adaptive voltage scaling scheme (AVS), a control loop is used to adjust the voltage level [20].

Schemes of finer granularity include voltage islands which are an extension of voltage scaling combined with the power gating technique, where blocks are switched-off if inactive [13]. Blocks are physically grouped to form islands with the same voltage levels and switching activities. This extension is useful for circuits with voltage scaling enabled at the block level, as the power distribution network (PDN) is simplified [21]. In addition, thermal issues are mitigated as power relates to heat dissipation [22].

2.3. Related Work for 3-D ICs

Several researchers [15], [23]-[28] have investigated the formation of voltage islands for 3-D ICs. This interest is due to the importance of voltage islands to issues relating to three-dimensional integration that require attention, such as thermal dissipation [17], power distribution network complexity [18], and process variations [19].

Zhu *et al.* explore several policies for task migration and DVFS to better manage heat in 3-D ICs [23]. Emphasis is placed at the micro-architecture level rather than the physical level. Xu *et al.* utilize a mixed integer linear programming (MILP) model for voltage-island generation optimizing the heat distribution and routing resources of the power distribution network, while maintaining the circuit performance in the three dimensions [24]. Moreover, a voltage assignment method for minimizing power in 3-D ICs is proposed in [25] while considering thermal and timing overheads due to level shifters. However, in both [24] and [25], the voltage levels of each block are assumed known. Furthermore, in these works the effect of the parasitic impedance of the TSVs on the power, performance, and voltage level of the 3-D blocks is not considered.

Additionally, Zhan *et al.* proposed a partition-based algorithm for assigning modules at the floorplan level to reuse currents between voltage domains and minimize the power dissipated on the ground distribution network [26]. Kapadia and Pasricha proposed a synthesis framework and a methodology to optimize the power distribution network in MVS mesh-based 3-D networks-on-chip (NoC) [27]. In both of these works, emphasis is placed on the effect of the power distribution network to the power of the 3-D NoC rather than how three-dimensional integration affects the operating voltage of a block/system considering the related speed and power.

Voltage assignment in 3-D ICs under process variations is considered in [15], [28]. Lee *et al.* proposed a grid-based multiple supply voltage method to statistically minimize power while considering spatially correlated process variations and thermal effects for 3-D ICs [15]. A first-order statistical timing model is utilized to capture the performance of the system based only on inter-block connections without considering the *RC* parasitic impedance of TSVs. This assumption limits the accuracy of this method as three-dimensional integration affects both intra-block and inter-block nets [3] and the voltage applied to a circuit can differ between the two- and three-dimensions. Finally, a post-silicon tuning methodology for improving the parametric yield of 3-D ICs with tier-adaptive-voltage-scaling is described in [28]. A few synthetic paths are assumed in this analysis and a generic statistical framework is utilized to evaluate the described approach rather than realistic benchmark circuits and accurate back-end timing information (STA-level) of the circuits.

Unlike prior works, in this paper the voltage requirements are not constrained to be the same between the 2-D and 3-D implementation of a circuit. This notion is based on the fact that the power and performance of a 3-D stacked circuit differ from the planar circuit at the same operating voltage. Hence, this situation provides a voltage headroom which can be utilized to enhance either power or speed in three-dimensional circuits.

3. Characteristics of TSV-based 3-D Circuits

In a multi-tier circuit, tiers are stacked and vertically interconnected with through silicon vias. The salient feature of this technology is the considerably shorter interconnect length [1]. Therefore, the power and/or the performance of 3-D stacked circuits differ from planar circuits at the same operating voltage. This situation is demonstrated in several works [1]-[5]. The reduction of the *RC* parasitic impedance in critical paths leads to an increased slack for 3-D circuits, thereby improving the performance. Alternatively, power reduction in 3-D ICs is achieved mainly due to the reduction of the total interconnect capacitance. The different origin of the power and speed enhancements constrains the available headroom for voltage scaling in the resulting 3-D ICs. Depending on how much the speed and power of the resulting 3-D system differ from the 2-D implementation at the same voltage, the voltage can be scaled accordingly to either enhance or compensate for any loss in these objectives.

However, three-dimensional integration is not beneficial for all circuits as excessive usage of TSVs can lead to increased wirelength [6]. In this case, the introduction of the third dimension degrades the power and performance of the circuit, not allowing the voltage to be scaled. In the case where three-dimensional integration provides some power savings, these gains can be utilized to increase the speed of the system by increasing the operating voltage while ensuring that $P_{3-D} = P_{2-D}$. This approach is applicable irrespective of whether the third dimension leads to an increase in speed. However, the moderate power savings in 3-D circuits from wirelength reduction and the quadratic relation of power to voltage heavily restrain the available voltage headroom. Therefore, considering the linear dependency of frequency to the dynamic power, the resulting increase in speed is marginal. In addition, the increase in voltage is limited due to the high thermal densities in 3-D circuits. Indeed the higher voltage levels lead to an increase in power and as a result in greater heat generation.

Alternatively, reducing the operating voltage by exploiting the additional timing slack that three-dimensional integration supports is a meaningful way to enhance the power savings in 3-D ICs. This method, exploiting the speed improvements in 3-D circuits, offers an alternative means to reduce power in 3-D

ICs compared to the traditional approach of the decreased interconnect capacitance. However, the applicability and the gains of this technique vary according to the available timing slack of the critical paths. This timing, in turn, is strongly dependent upon the characteristics of the gates and wires composing these paths. A methodology, which considers the timing behavior of the paths, for identifying which circuits can benefit from this approach is presented in the next section.

4. Determine whether voltage reduction is applicable

As discussed in the previous section, voltage scaling is not applicable to all 3-D ICs. In addition, evaluating the speed and power tradeoff in these circuits by only considering the inter-tier paths is not sufficient to determine voltage scaling as demonstrated by the results (see Section 6). Consequently, the critical paths of a circuit should be carefully examined to early evaluate whether voltage reduction degrades the target performance. Therefore, an enhanced timing model for circuit paths is formulated in Subsection 4.1, including several parameters, such as the number and type of gates, the interconnect segments within the paths, and the gate delay sensitivity to voltage. In Subsection 4.2, this model is incorporated into a methodology to determine when circuits can exploit voltage reduction in three dimensions.

4.1. Interconnect and voltage aware timing model

The delay of a logic path depends both upon the number and type of gates and the wires interconnecting these gates. The method of logical effort [29] is a useful technique for estimating the gate delay in CMOS circuits. Based on the extension of this method in [30], where interconnects are also considered, the delay of a N-stage path (see Fig. 1), normalized to the delay of a minimum sized inverter (τ), is

$$d_{path} = \sum_{i=1}^{N} \Big(g_i \cdot (h_i + h_{w_i}) + (p_i + p_{w_i}) \Big), \tag{4}$$

where g_i and p_i are, respectively, the logical effort and parasitic delay related to the characteristics of the logic gate. These parameters for different gates considering simple layout styles are obtained from [29]. The parameter h_i is the electrical effort of the gate defined as the ratio of input capacitance of two successively connected gates (C_{i+1}/C_i), as depicted in Fig. 1 for gates g_i and g_{i+1}. The capacitive effort h_{w_i} and the resistive effort p_{w_i} for interconnects are, respectively [30],

$$h_{w_i} = l_i \cdot c_{w_i}/C_i, \tag{5}$$

$$p_{w_i} = l_i \cdot r_{w_i} \cdot (0.5 \cdot l_i \cdot c_{w_i} + C_{i+1})/\tau, \tag{6}$$

where l_i is the length of the wire and r_{w_i}, c_{w_i} are, respectively, the resistance and capacitance per unit length depending whether the wire is composed of local, intermediate, and/or global metal layers. However, in this model the sensitivity of the gate delay to voltage is not considered.

The effects of voltage and temperature variations on logical effort are modeled in [31] for several process nodes and for different operating regions of MOSFETs, such as strong, moderate and weak inversion. Considering that transistors in digital circuits typically operate in strong inversion, the logical effort is

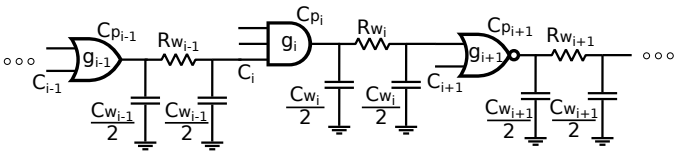

Fig. 1. A typical path composed of gates and interconnects.

$$g_i' = g_i \cdot g_u, \tag{7}$$

$$g_u = \frac{V_{dd}}{A(T) \cdot (V_{dd} - V_{th0} + k \cdot T)^{(3/2)}}, \tag{8}$$

where g_u is a fitting function obtained from [31] to capture the effect of voltage and temperature on logical effort across different technologies. $A(T)$ is a second order polynomial function of temperature (T) depending upon the technology, V_{th0} is the voltage threshold at $0°C$ and k is the slope of V_{th} as a function of temperature. At the 45 nm technology node, for $V_{th0} = 0.46V$, $k = 6.32 \cdot 10^{-4}$, and $T = 25°C$ [31]

$$g_u = \frac{V_{dd}}{2.413 \cdot (V_{dd} - 0.4442)^{(3/2)}}, \tag{9}$$

where for $V_{dd} = 1V$, $g_u = 1$. Therefore, by rewriting (4) to include the voltage sensitivity, the delay of the path is

$$d_{path} = \sum_{i=1}^{N} \Big(g_i \cdot g_u \cdot (h_i + h_{w_i}) + (p_i + p_{w_i}) \Big). \tag{10}$$

With this model, the delay of a critical path CP is described as a function of the number (N) and type of gates (g_i, C_i), length of the interconnect segments (l_i) and operating voltage (V_{dd})

$$d_{path_{CP}} = f(N, g_i, C_i, l_i, V_{dd}). \tag{11}$$

4.2. Timing-slack voltage reduction methodology for 3-D ICs

A standard design flow (assuming this flow can incorporate the third dimension) can provide all of the timing information across different operating voltages for a 3-D circuit and determine any useful voltage reduction. However, this process is highly timing consuming. Therefore, several guidelines are presented in this subsection for identifying early in the design process, which two-dimensional circuits allow for voltage reduction with vertical integration.

The objective of this analysis is to roughly estimate the slack of the paths in a 3-D circuit (e.g. $\Delta d_{2Dto3D} = d_{2D} - d_{3D}$) at the same operating voltage and utilize any increased slack (i.e. $\Delta d_{2Dto3D} > 0$) for voltage reduction such that $\Delta d_{2Dto3D} = 0$ at $V_{dd_{3D}} < V_{dd_{2D}}$. In order to identify if voltage reduction is possible for a circuit, the critical paths of a 2-D circuit are grouped and sorted in ascending order of slack (i.e., descending order of criticality). The post-routing timing information should be considered for these critical paths such that the RC parasitic

impedance of the wires is included in the delay of the path. Afterwards, a user-defined threshold is applied to select the number of paths which will be analyzed. This threshold is effectively a knob trading off the accuracy in projecting the available slack and voltage reduction in 3-D ICs with computational time. Using few paths leads to a fast analysis. However, the accuracy is low, as some non-critical paths can exhibit great increase in delay with voltage reduction and, therefore, become critical. Hence, representative paths with various slacks and logic depths should be selected and analyzed. The guidelines in this section facilitate this process.

A N-stage path comprises N_{wires} segments with various lengths. These segments are grouped based on the metal layers used for the routing of each segment as

$$N_{wires} = N_{local} + N_{int} + N_{global}, \tag{12}$$

where N_{local}, N_{int}, and N_{global} is the number of wire segments laid out within the local, intermediate, and global metal layers, respectively. To determine how the length of the wire segments is affected by the third dimension, a function which describes the length of wires (l_i) in two and three dimensional circuits ($l_{i_{3D}} = f(l_{i_{2D}})$) is required. This function can be obtained from wirelength prediction models, such as [6], [32]. The delay of the critical paths in the 3-D circuit is projected by incorporating this function with the model described in Section 4.1. The integration of the wirelength and timing models intends to provide a pass or fail check, indicating whether wirelength reduction within the paths leads to useful voltage reduction with the introduction of the third dimension.

This model is applied to selected critical paths according to the chosen threshold. Without loss of generality, this analysis is carried out on a 45 nm technology, where the input capacitance of the gates is obtained from [33]. For different technology nodes, the function of g_u can be fitted similarly to [31] and the input capacitance is obtained from the corresponding technology libraries. Moreover, a first-order wirelength model for 3-D circuits is used [34]

$$l_{i_{3D}} = \frac{l_{i_{2D}}}{\sqrt{n}}, \tag{13}$$

where n is the number of tiers of the 3-D stack. This function is applied to intermediate and global wires, whereas the length of the local interconnect segments is assumed to not change in the 3-D stack [34]. The inherent inaccuracy of wirelength models do not permit a highly accurate estimate for the additional slack. Rather the precise increase in slack is determined by a complete design flow (see Section 5). In general, (13) offers a loose upper bound for describing the decrease in wirelength as the overhead of TSV area in wirelength is not captured, however this model offers useful estimates as demonstrated by the results in Section 6.

As observed by many interconnect prediction models, three-dimensional integration reduces the wirelength of a circuit by decreasing the length of long intermediate and global interconnects [1]. If these long wires belong to the critical paths of the circuit, the additional timing slack from the *RC* reduction can be utilized to reduce the voltage. Guideline 1, *circuits where critical paths comprise only local interconnects are not expected to support voltage reduction in three dimensions.* In addition, wirelength reduction in 3-D ICs depends on several factors, such as the size and number of the TSVs, and the number of tiers [6]. Therefore, at the early design stages of a 3-D circuit, the wirelength prediction models should be employed to determine whether wirelength decreases for the target 3-D technology. Guideline 2, *when the critical paths of a 2-D circuit comprise long intermediate and global wires and wirelength reduction is predicted for the 3-D technology, voltage reduction is applicable to this circuit.*

Applying (10) to the critical paths of a planar circuit and the projected wirelength reduction within these paths from (13), an estimate for the added timing slack in a 3-D circuit is determined. The projected decrease in the delay of several paths based on this model for the benchmark circuits in Section 6 is depicted in Fig. 2. The increased timing slack from the introduction of the third dimension depends upon both the number of the gates and the length of the wire segments of the path. Consequently, as shown in this figure paths with long global wires exhibit greater delay reduction as compared to paths with shorter intermediate wires due to the larger wirelength reduction from vertical integration. In addi-tion, less gate-dominated paths exhibit a larger speedup than paths with more gates for the same wire *RC* reduction. Guide-line 3, *the higher the portion of the interconnect delay in a path, the greater the increase in the timing slack due to the vertical integration and thereby the larger the decrease in voltage.*

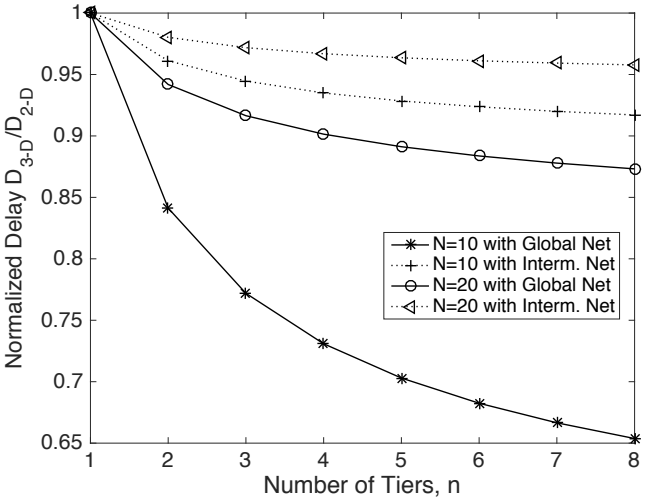

Fig. 2. Delay reduction due to vertical integration for paths with different number of gates (N) and the same global and intermediate wire segments, where wirelength changes with n according to (13).

A tradeoff, however, exists between the added timing slack in the 3-D ICs and the increase in the delay due to voltage reduction. Considering (10), the rate of change in delay depends linearly upon the interconnect length ($\frac{\partial d_{path}}{\partial l} \propto l$) and superlinearly to the operating voltage ($\frac{\partial d_{path}}{\partial V_{dd}} \propto -\frac{1}{V_{dd}^{3/2}}$). This behavior indicates that the gates comprising the paths of the circuit primarily determine the capability to scale voltage as compared

to wirelength reduction. Consequently, the effect of decreasing the voltage on the delay of the gates should also be considered to determine the sheer advantage of voltage scaling. This requirement has driven the development of a complete RTL flow described in Section 5 to holistically and accurately determine the timing slack in a 3-D circuit.

The sensitivity of the delay of a gate to voltage reduction is

$$s_{gate_i} = \left| \frac{\Delta d_{gate_i}}{\Delta V_{dd}} \right|, \tag{14}$$

where Δd_{gate_i} is the delay increase of the gate i for voltage reduction of ΔV_{dd}. Gates with lower sensitivity utilize more efficiently a given amount of voltage reduction, in terms of less added delay, as compared to gates with higher sensitivity. The delay of diverse gates with different driving strengths exhibits different sensitivity to voltage, as depicted in Fig. 3. This delay sensitivity to voltage (14) is described by a first-order model using the first partial derivative of (10) to V_{dd} for a given gate i,

$$s_{gate_i} = \left| \frac{\partial d_{gate_i}}{\partial V_{dd}} \right| = \frac{g_i}{C_i} \cdot C_{load_i} \cdot \left| \frac{\partial gu}{\partial V_{dd}} \right|, \tag{15}$$

$$C_{load_i} = C_{i+1} + l_i \cdot C_{w_i}, \tag{16}$$

where $\frac{\partial gu}{\partial V_{dd}}$ is technology dependent and C_{load_i} is the load capacitance that the gate i drives. The sensitivity of the delay of a gate to the changing voltage depends upon the factor $\frac{g_i}{C_i}$, as the load capacitance is not a characteristic of the gate. The delay of stronger gates (higher C_i) with low logical effort (g_i) is less sensitive to voltage changes as compared to weaker and higher logical effort gates.

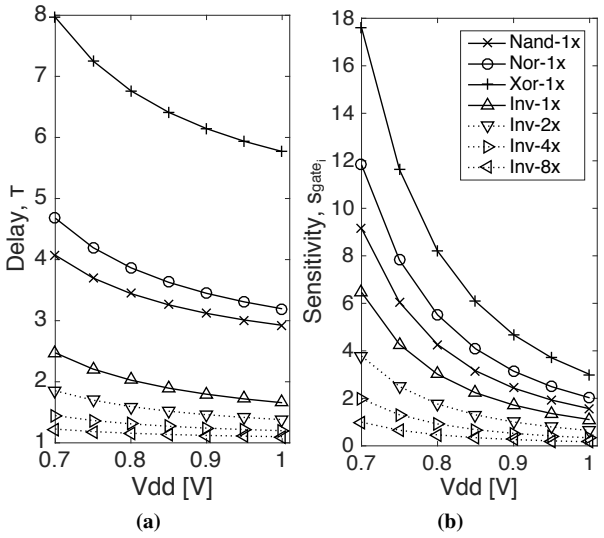

Fig. 3. Delay (a) and sensitivity (b) of logic gates to voltage reduction while driving a minimum size inverter.

As a path is composed of both gates and wires, the sensitivity of a path is described by adding the product of the sensitivity of each gate and the load capacitance,

$$s_{path} = \left| \frac{\partial d_{path}}{\partial V_{dd}} \right| = \sum_{i=1}^{N} \frac{g_i}{C_i} \cdot C_{load_i} \cdot \left| \frac{\partial gu}{\partial V_{dd}} \right|. \tag{17}$$

This equation indicates that the potentially reduced interconnect capacitance due to vertical integration decreases the delay sensitivity of the paths as C_{load} decreases. Hence, the delay of a 3-D circuit can be less sensitive to voltage reduction as compared to the 2-D counterpart. For two critical paths with different sensitivities to voltage (taken from the benchmark circuits in Section 6), the projected voltage reduction is depicted in Fig. 4. Both paths are assumed to exhibit the same decrease in wirelength.

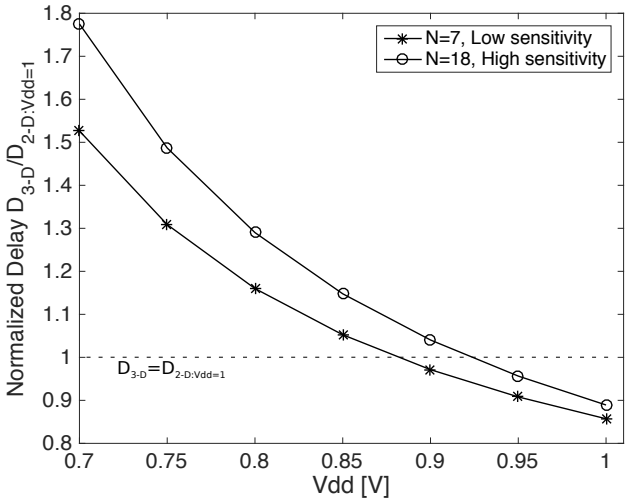

Fig. 4. Change in delay of paths with different path sensitivity where voltage is gradually reduced.

For the path with high sensitivity to voltage, the additional slack from the third dimension is absorbed fast by the increase in delay of the many gates ($N = 18$), resulting in only 7% voltage decrease. Alternatively, the path with the low sensitivity to voltage, which comprises fewer ($N = 7$) and less sensitive gates, absorbs the additional slack at a slower rate with the change in voltage. This situation leads to an almost double reduction in voltage (13%) without affecting the original delay of the 2-D path ($D_{3-D} = D_{2-D:V_{dd}=1}$). Guideline 4, *for 3-D circuits with paths that exhibit low sensitivity to voltage (according to (17)) and long wires, voltage reduction is higher.*

5. Design flow for voltage reduction in 3-D ICs

If the application of the model presented in Section 4 indicates an increased slack (i.e. $\Delta d_{2Dto3D} > 0$) and an acceptable voltage reduction from the third dimension, a methodology to precisely determine the allowed decrease in voltage and the related power savings is required. In addition, the effect of different TSV technologies and bonding styles on timing should be considered. An EDA compatible flow for applying and quantifying the savings in power from voltage reduction in 3-D stacks, is described in this section.

As shown in Fig. 5, most of the steps are based on existing commercial tools. The inputs to this flow are the Verilog/VHDL description of the design and the same timing constraints and

operating voltage as in the two-dimensional circuit since iso-performance operation is a requisite. The first step produces a synthesized netlist which is input to the 3-D floorplanner, generating a floorplan for each tier based on the selected TSV technology and bonding style. The circuit partition among the tiers and the position of the TSVs are determined by the 3-D Craft [35]. Moreover, a commercial tool, such as Encounter [36], is utilized to place and route the cells in each tier. After these steps, the netlists from each tier are merged while performing a design equivalence check through the Formality tool [37]. In addition, the standard-parasitic-exchange-format (SPEF) file of each tier are merged into a global SPEF adding the parasitic impedance of TSVs, as described in [38]. In the last step, a timing analysis tool, such as PrimeTime [39], is utilized to determine the new and increased slack.

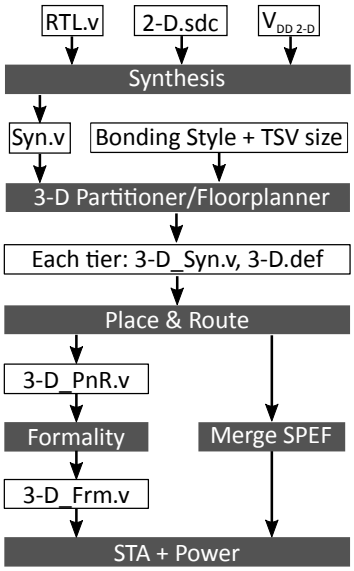

Fig. 5. Stages of a 3-D design flow where commercial EDA tools with standard file formats are utilized.

The methodology for evaluating voltage scaling in 3-D circuits is depicted in Fig. 6. Initially, the timing of the 3-D circuit is analyzed by performing STA at the 3-D stack for the same operating voltage as in the 2-D circuit ($Vdd_{3-D} = Vdd_{2-D}$). During this step, the exact gained (or lost) slack from the usage of the third dimension is obtained. In the case where the delay of the 3-D circuit is increased, voltage reduction is prohibited. However, following the guidelines and using the timing model in Section 4, this situation can be avoided. If no additional slack is predicted by the model, smaller TSVs or different bonding styles need to be considered during partitioning to minimize the overhead of TSVs on the speed of the circuit.

Alternatively, when three-dimensional integration enhances the performance of the circuit, the increased slack is utilized for voltage reduction in the 3-D stack. To perform timing analysis across multiple voltage levels, interpolation between different process, voltage, and temperature (PVT) libraries is required [39]. Therefore, for a range of operating voltages and by starting from the same operating voltage as in the 2-D circuit, voltage is gradually reduced ($Vdd_{3-D} < Vdd_{2-D}$) and STA is

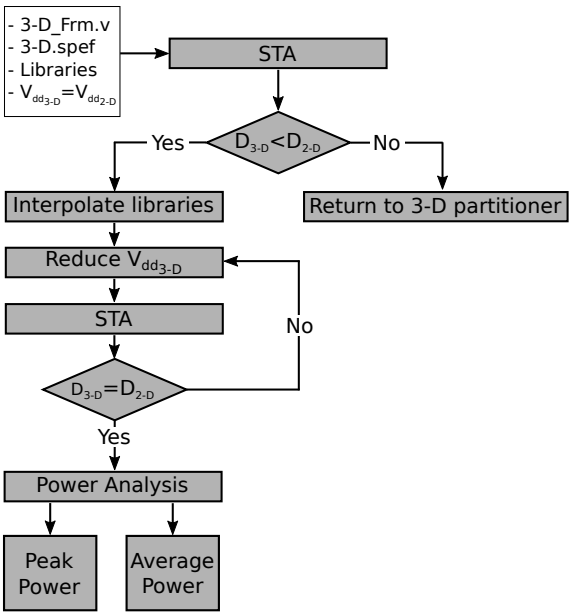

Fig. 6. Methodology to evaluate the voltage reduction in 3-D circuits.

performed. During this voltage sweep, the same process and temperature variations are assumed. The minimum operating voltage for the 3-D circuit is reached when the performance is equivalent to the original 2-D circuit ($D_{3-D} = D_{2-D}$). Afterwards, power analysis is performed to quantify the power improvements of the 3-D stack as compared to the 2-D circuit. Average power analysis, assuming toggling rates for nets and cells, is performed for battery life considerations. Moreover, cycle-accurate events produced from testbenches during back-annotated simulation of the circuits are used to evaluate the peak power of the 3-D stack.

6. Results

The voltage scaling methodology is applied to several benchmark circuits implemented in three dimensions in Section 6.1. The power savings from this approach are quantified in Section 6.2. PrimeTime [39] is used for timing and both average and time-based power analysis.

6.1. Applicability of Voltage scaling to 3-D ICs

Several benchmark circuits are utilized to evaluate any potential gain in power due to the change in voltage enabled by the reduced wirelength. The benchmark circuits are listed in Table 1. These circuits are evaluated at a 45 nm and a 65 nm technology from TSMC [33]. As advanced technologies are utilized, via-first TSVs with a diameter 1 μm, length 10 μm, resistance 334 $m\Omega$, and capacitance 3 fF are assumed for the vertical interconnects [40], [41]. Moreover, two tiers bonded face-to-back are assumed.

The characteristics of the benchmark circuits in both the two- and three-dimensions are reported in Table 2. According to these results, the vertical integration improves the wirelenth for several circuits at the 45 nm technology. Due to the

Table 1
Benchmark circuits.

Circuits	Reference
B04, B19, B20	[42]
LDPC, AES, DES3, FFT	[43]
AVA	[44]

very small size of B04 circuit, any insertion of TSV increases considerably the area resulting in increased wirelength. The large number of TSVs for the LDPC circuit in addition to the small technology node lead to increased wirelength, although the speed of this circuit is dominated by the delay of the wires. However, this situation changes when utilizing an older technology node (65 nm).

Table 2
Area, wirelength, and number of TSVs for the benchmark circuits designed both in two and three dimensions.

Circuits	Tech. node	# of Cells [K]	2-D Area [μm^2]	2-D WL [mm]	3-D Area [μm^2]	3-D WL [mm]	3-D # of TSVs
B04	45	0.31	1.038	3.4	1,315	4.3	123
B19	45	66.1	155,304	919	166,502	795	745
B20	45	12.1	22,508	130	23,552	105	545
AVA	45	12.3	38,575	257	40,528	215	651
LDPC	45	40.0	107,605	1,974	113,258	2,325	3,820
LDPC	65 LP	67.0	469,593	6,328	479,770	4,645	4,384
DES3	45	50.7	102,749	666	106,923	615	925
AES	45	117.3	232,545	2,330	237,998	2,211	934
FFT	45	242.7	713,982	4,656	718,365	4,067	995

In order to reduce the operating voltage, a circuit must exhibit an increase in speed in three dimensions as compared to the planar implementation of the circuit at the same operating voltage. The early analysis of the critical paths of a planar circuit, as discussed in Section 4, determines whether voltage reduction is possible for the target 3-D technology. Therefore, representative paths, the slack of which is up to 20% of the clock period, are examined for the benchmark circuits, as described in Section 4. In addition, the 3-D EDA flow described in Section 5 (see Fig. 5), is utilized to accurately quantify the additional timing slack of the circuit in three-dimensions. The resulting slack for the benchmark circuits in the two-tier stack from the application of the EDA flow and the timing model of Section 4, respectively, are listed in Table 3. Note, that only for the bench-mark circuits which exhibit wirelength reduction, performance is reported in this table.

Based on the proposed model B20 and AES fail to show any speed improvement in three dimensions. This behavior, is due to the fact that the speed of these circuits is primarily determined by the delay of the gates (gate-dominated circuit), where only local wire segments are utilized to interconnect the gates within the critical paths. Considering guideline 1 and (13), no added timing slack is predicted by our model and reducing the voltage is not an option for these circuits. Indeed, the application of the flow to B20 and AES yields a negative slack due to the overhead of the TSVs which negatively affects the delay.

Table 3
Supported clock period of the benchmark circuits for the same operating voltage as in the 2-D design.

Circuits	Tech. node	V_{dd} [V]	2-D T_{clk} [ps]	3-D T_{clk} [ps]	Slack [ps]	Diff. [%]	Our Model
B20	45	1	1,125	1,171	-46	-4.1	(-) Fail
B19	45	1	1,273	1217	56	4.4	(3.6%) Pass
AVA	45	1	1,522	1,304	218	14.3	(11%) Pass
LDPC	65 LP	1.2	4,278	3,468	810	18.9	(22%) Pass
DES3	45	1	1,068	1,043	25	2.4	(1.8%) Pass
AES	45	1	946	1,152	-206	-21.7	(-) Fail
FFT	45	1	1,421	1,205	216	15.2	(14.9%) Pass

Moreover, for circuits such as B19 and DES3, where a small portion of the delay is due to some short intermediate wires within the critical paths (see guideline 3), our model predicts a positive slack of 3.6% and 1.8%, respectively. A marginal improvement is also determined by the flow described in Section 5. As three-dimensional integration reduces slightly the delay of these circuits, voltage reduction is limited.

Alternatively, wire-dominated circuits, such as AVA, LDPC, and FFT, exhibit great performance improvements of 14.3%, 18.9% and 15.2%, respectively, as compared to the 2-D design at the same operating voltage. This situation is also predicted by our timing model and guideline 2. Note that for the LDPC circuit, the longest inter-tier path exhibits 31% delay reduction. However, the clock period is constrained by paths within a tier (intra-tier paths) and as a result the delay decreases only by 18.9%. This example demonstrates that considering only the inter-tier paths is not sufficient to determine the actual increase in slack and thereby the potential voltage reduction in 3-D ICs.

The increased slack from the introduction of the third dimension is exploited to reduce the voltage by utilizing the flow in Section 5 (see Fig. 6). The minimum operating voltage for the benchmark circuits at iso-performance operation with the 2-D circuits is listed in Table 4. The operating voltage reduction for the circuits B19 and DES3 is rather negligible as the increased slack from the vertical integration is limited. Alternatively, considerable voltage decrease is obtained for the rest of the circuits. The greatest voltage reduction is obtained for the LDPC circuit, where global long wires are utilized in the critical paths of the planar implementation of the circuit.

Table 4
Reduction of the operating voltage for the benchmark circuits due to the added timing slack produced by the 3-D stacking.

Circuits	2-D V_{dd} [V]	3-D V_{dd} [V]	Reduction [%]
B19	1.0	0.98	2
AVA	1.0	0.92	8
LDPC	1.2	1.03	14
DES3	1.0	0.97	3
FFT	1.0	0.88	12

Interestingly, for almost the same timing slack (~217 ps) for AVA and FFT, the operating voltage for the FFT circuit is decreased up to 12% whereas for the AVA this decrease is only 8%. Considering guideline 4, the FFT circuit exhibits greater voltage reduction as the delay of the critical paths is less sensitive to voltage reduction than in the AVA circuit. This behavior is predicted from our model in Fig. 4, where the critical path of FFT ($N = 7$) with low sensitivity to voltage, absorbs the additional slack at a slower rate with the change in voltage than the critical path of AVA ($N = 18$). This situation leads to a greater voltage reduction for the FFT in order to meet the delay of the 2-D circuit.

6.2. Quantifying Power Gains

In this subsection, the power consumption of the benchmark circuits where voltage reduction is applicable to vertical integration is investigated. The flow illustrated in Fig. 6 is utilized to measure the power of the circuits under different scenarios. The simulation scenarios are listed in Table 5. The baseline scenario (S1) is the 2-D circuit for a specific frequency and operating voltage. Scenario S2 is selected to quantify the power savings from 3-D integration only due to the decreased interconnect capacitance. Finally, our approach for reducing the operating voltage while keeping the same operating frequency as in the 2-D implementation is considered in scenario S3. The resulting operating voltage for each circuit is listed in Table 4.

Table 5
Simulated scenarios for evaluating power consumption.

Scenarios	Integration	Frequency	Operating Voltage
S1	2-D	F_{2-D}	V_{2-D}
S2	3-D	$F_{3-D} = F_{2-D}$	$V_{3-D} = V_{2-D}$
S3	3-D	$F_{3-D} = F_{2-D}$	$V_{3-D} < V_{2-D}$

For the average power analysis, the toggle rate of all cells and nets within each circuit is set to 20%. The total average power consumed by the circuits in all of the scenarios is illustrated in Fig. 7. In addition, the breakdown of the total power into the different power components is listed in Table 6. The decrease of interconnect capacitance in 3-D ICs results, on average, in 6.6% power reduction for the benchmark circuits. The power savings of this approach are limited, as the power dissipated by the cells is not significantly reduced. Alternatively, our approach of reducing the voltage while meeting the 2-D timing constraints leads on average to 22.3% power reduction, as this approach exploits the quadratic relation of voltage to the dynamic power of both cells and nets.

Furthermore, application-specific testbenches from [43] - [44] are utilized to quantify the total (cycle-accurate) and peak power consumption. These testbenches simulate real-task events for the benchmark circuits, such as decoding messages for LDPC and AVA, encrypting and decrypting messages for DES3, and fast fourier transformation of a signal for the FFT. In Fig. 8 the power trace for the LDPC circuit is plotted for the scenarios listed in Table 5. Three-dimensional integration reduces the

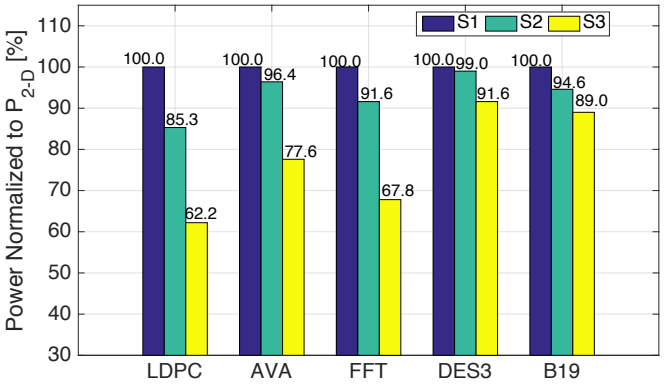

Fig. 7. Total average power consumption at the same speed $D_{2-D} = D_{3-D}$.

Table 6
Breakdown of average power to its components.

Circuit	Scenario	Nets $[mW]$	Cells $[mW]$	Leakage $[\mu W]$
LDPC	S1	20.49	8.76	0.024
	S2	16.26	8.69	0.024
	S3	11.88	6.31	0.015
AVA	S1	0.51	0.47	1.07
	S2	0.48	0.46	1.07
	S3	0.40	0.36	0.67
FFT	S1	10.93	16.15	25.3
	S2	8.18	16.12	25.3
	S3	6.16	11.84	13.1
DES3	S1	1.58	2.12	3.5
	S2	1.55	2.11	3.5
	S3	1.44	1.95	2.9
B19	S1	2.22	2.11	4.85
	S2	2.02	2.09	4.85
	S3	1.91	1.95	4.23

power by 10% due to shorter wires, while our approach of reducing the voltage achieves a 34% decrease in power. In addition, with the proposed approach the peak power is reduced by 27% as compared to the 3-D implementation where the voltage is not changed ($V_{3-D} = V_{2-D}$).

The power savings of application-specific testbenches for the investigated circuits are depicted in Fig. 9. In addition, the breakdown of the total power consumption and the peak power for executing real tasks on circuits are listed in Table 7. AVA and DES3 exhibit a higher total and peak power as compared to two dimensions. This behavior is due to the small percentage of switching wires during these tasks and these nets contain TSVs. This example demonstrates the limitations of 3-D integration to provide power reduction if only the reduction in the wire capacitance is considered as has been done by several previous works. For the FFT circuit, reducing the voltage results in 28.5% and 29.8% decrease in total and peak power, respectively, as compared to the 3-D case where $V_{3-D} = V_{2-D}$ (scenario S2). This situation demonstrates the effectiveness of the proposed approach to significantly decrease power

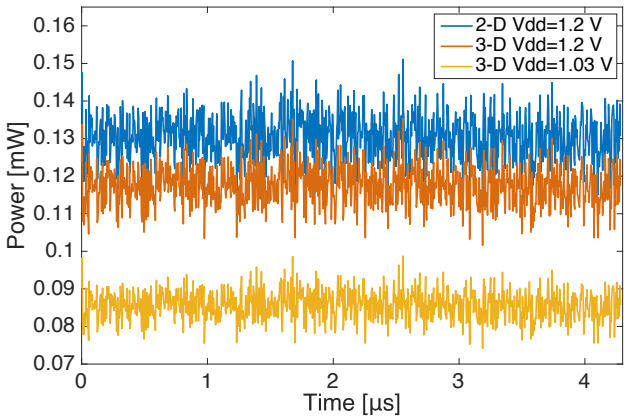

Fig. 8. Power trace of the LDPC circuit for different scenarios.

without compromising performance. Furthermore, the specific characteristics of the circuit are captured accurately predicting whether voltage reduction at iso-performance is an option for this circuit in three-dimensions.

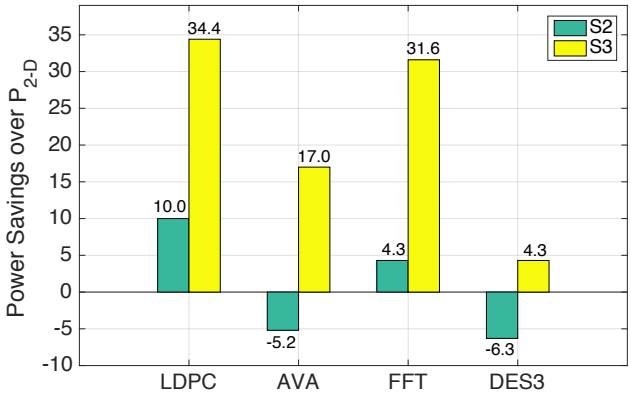

Fig. 9. Power savings of application-specific testbenches.

Table 7
Breakdown of total (cycle-accurate) power to its components and peak power for application-specific testbenches.

Circuit	Scenario	Nets [μW]	Cells [μW]	Leakage [μW]	Peak [μW]
LDPC	S1	84.9	45.5	0.024	151.1
	S2	72.0	45.3	0.024	135.3
	S3	52.6	33.0	0.015	98.7
AVA	S1	1.6	9.8	1.14	16.1
	S2	2.3	9.8	1.14	16.8
	S3	1.9	7.9	0.71	13.2
FFT	S1	32.9	101.7	25.2	181.9
	S2	26.4	101.3	25.2	178.2
	S3	20.5	76.0	13.0	125.1
DES3	S1	2.8	12.2	3.5	20.7
	S2	3.9	12.4	3.5	22.0
	S3	3.6	11.3	2.9	19.9

7. Conclusions

In this work, voltage scaling in TSV-based three-dimensional circuits is investigated. The objective is to exploit the additional slack on critical paths from the introduction of the third dimension to enhance power savings by reducing the supply voltage. Several guidelines and a timing model based on the logical effort are offered for identifying early in the design process if 2-D circuits can benefit from voltage reduction with vertical integration. In addition to this qualitative model, a methodology for applying voltage reduction and quantifying the power savings in 3-D ICs is presented. The traditional notion where the power is reduced due to the decrease in the wire capacitance of a 3-D IC leads to low and often inadequate power savings for several benchmark circuits (6.6 % on average). Alternatively, the proposed approach results in power reduction of 22.3% on average. Moreover, a reduction of 27% in the peak power is observed for an LDPC circuit as compared to the 3-D case where the nominal voltage is not changed and the same speed as in two-dimensions is maintained.

References

[1] V. F. Pavlidis and E. G. Friedman, *Three-dimensional Integrated Circuit Design*, Morgan Kaufmann Publishers, 2009.
[2] B. Vaidyanathan *et al.*, "Architecting Microprocessor Components in 3-D Design Space," *Proceedings of the IEEE International Conference on VLSI Design*, pp. 103-108, January 2007.
[3] B. Black *et al.*, "Die Stacking (3D) Microarchitecture," *Proceedings of the IEEE/ACM International Symposium on Microarchitecture*, pp. 469-479, December 2006.
[4] K. Puttaswamy and G. H. Loh, "3D-Integrated SRAM Components for High-Performance Microprocessors," *IEEE Transactions on Computers*, Vol. 58, No. 10, pp. 1369-1381, October 2009.
[5] Y.-F. Tsai *et al.*, "Design Space Exploration for 3-D Cache," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 16, No. 4, pp. 444-455, April 2008.
[6] D. H. Kim, S. Mukhopadhyay, and S. K. Lim, "TSV-Aware Interconnect Distribution Models for Prediction of Delay and Power Consumption of 3-D Stacked ICs," *IEEE on Computer-Aided Design of Integrated Circuits and Systems,* Vol. 33, No. 9, pp. 1384-1395, September 2014.
[7] J. W. Joyner *et al.*, "A Three-Dimensional Stochastic Wire-Length Distribution for Variable Separation of Strata," *Proceedings of the IEEE International Interconnect Technology Conference,* pp. 126-128, June 2000.
[8] O. Coudert, "Gate Sizing for Constrained Delay/Power/Area Optimization," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems,* Vol. 5, No. 4, pp. 465-472, December 1997.
[9] W. Qing, M. Pedram, and W. Xunwei, "Clock-Gating and its Application to Low Power Design of Sequential Circuits," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications,* Vol. 47, No. 3, pp. 415-420, March 2000.
[10] J. Hailin, M. Marek-Sadowska, and S. R. Nassif, "Benefits and Costs of Power-Gating Technique," *Proceedings of the International Conference on Computer Design,* pp. 559-566, October 2005.
[11] M. Anis, S. Areibi, and M. Elmasry, "Design and Optimization of Multithreshold CMOS (MTCMOS) Circuits," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* Vol. 22, No. 10, pp. 1324-1342, Oct. 2003.
[12] M. Keating *et al.*, *Low Power Methodology Manual: For System-On-Chip Design,* ISBN: 9781441944184, Springer, 2007.
[13] D. E. Lackey *et al.*, "Managing Power and Performance for System-On-Chip Designs using Voltage Islands," *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design,* pp. 195-202, November 2002.

[14] D. H. Kim *et al.*, "Design and Analysis of 3D-MAPS (3D Massively Parallel Processor with Stacked Memory)," *IEEE Transactions on Computers,* Vol. 64, No. 1, pp. 112-125, January 2015.
[15] S.-A Yu, P.-Y. Huang, and Y.-M. Lee, "A Multiple Supply Voltage Based Power Reduction Method in 3-D ICs Considering Process Variations and Thermal Effects," *Proceedings of the IEEE Asia and South Pacific Design Automation Conference,* pp. 55-60, January 2009.
[16] A. Das *et al.*, "Evaluating Voltage Islands in CMPs under Process Variations," *Proceedings of the International Conference on Computer Design*, pp. 129-136, October 2007.
[17] H. Xu, V. F. Pavlidis, and G. De Micheli, "Analytical Heat Transfer Model for Thermal Through-Silicon Vias," *Proceedings of the Conference on Design, Automation, and Test, in Europe,* pp. 395-400, March 2011.
[18] A. Todri-Sanial *et al.*, "Globally Constrained Locally Optimized 3-D Power Delivery Networks," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems,* Vol. 22, No. 10, pp. 2131-2144, October 2014.
[19] H. Xu *et al.*, "Timing Uncertainty in 3-D Clock Trees Due to Process Variations and Power Supply Noise," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems,* Vol. 21, No. 12, pp. 2226-2239, December 2013.
[20] T. Kuroda *et al.*, "Variable Supply-Voltage Scheme for Low-Power High-Speed CMOS Digital Design," *IEEE Transactions on Solid-State Circuits,* Vol. 33, No. 3, pp. 454-462, March 1998.
[21] B. Amelifard and M. Pedram, "Optimal Design of the Power-Delivery Network for Multiple Voltage-Island System-on-Chips," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* Vol. 28, No. 6, pp. 888-900, June 2009.
[22] W. L. Hung *et al.*, "Temperature-Aware Voltage Islands Architecting in System-on-Chip Design," *Proceedings of the IEEE International Conference on Computer Design,* pp. 689-694, October 2005.
[23] C. Zhu *et al.*, "Three-Dimensional Chip-Multiprocessor Run-Time Thermal Management," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* Vol. 27, No. 8, pp. 1479-1492, August 2008.
[24] N. Xu *et al.*, "Thermal-Aware Post Layout Voltage-Island Generation for 3D ICs," *Journal of Computer Science and Technology,* Vol. 28, No. 4, pp. 671-681, July 2013.
[25] S.-H. Whi and Y.-M. Lee, "Supply Voltage Assignment for Power Reduction in 3D ICs Considering Thermal Effect and Level Shifter Budget," *Proceedings of the IEEE International Symposium on VLSI Design, Automation and Test,* pp. 1-4, April 2011.
[26] Y. Zhan *et al.*, "Module Assignment for Pin-Limited Designs under the Stacked-Vdd Paradigm," *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design,* pp. 656-659, November 2007.
[27] N. Kapadia and S. Pasricha, "A Co-Synthesis Methodology for Power Delivery and Data Interconnection Networks in 3D ICs," *Proceedings of the IEEE International Symposium on Quality Electronic Design,* pp. 73-79, March 2013.
[28] K. Chae and S. Mukhopadhyay, "Tier-Adaptive-Voltage-Scaling (TAVS): A Methodology for Post-Silicon Tuning of 3D ICs," *Proceedings of the IEEE Asia and South Pacific Design Automation Conference,* pp. 277-282, January 2012.
[29] I. Sutherland, R. F. Sproull, and D. Harris, *Logical Effort: Designing Fast CMOS Circuits*, Morgan Kaufmann Publishers, 1999.
[30] A. Morgenshtein *et al.*, "Unified Logical Effort: A Method for Delay Evaluation and Minimization in Logic Paths With *RC* Interconnect," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 18, No. 5, pp. 689-696, May 2010.
[31] M.-H Chang *et al.*, "Logical Effort Models with Voltage and Temperature Extensions in Super-/Near-/Sub-Threshold Regions," *Proceedings of the IEEE International VLSI Design, Automation and Test Symposium*, pp. 1-4, April 2011.
[32] J. W. Joyner *et al.*, "A Three-Dimensional Stochastic Wire-Length Distribution for Variable Separation of Strata," *Proceedings of the IEEE International Interconnect Technology Conference,* pp. 126-128, June 2000.
[33] Taiwan Semiconductor Manufacturing Company, *www.tsmc.com*.
[34] M. Bamal *et al.*, "Performance Comparison of Interconnect Technology and Architecture Options for Deep Submicron Technology Nodes," *Proceedings of the International Interconnect Technology Conference*, pp. 202-204, May 2006.
[35] J. Cong and G. Luo, "A Multilevel Analytical Placement for 3D ICs," *Proceedings of the IEEE Asia and South Pacific Design Automation Conference,* pp. 361-366, January 2009.
[36] Cadence® Encounter® Version v13.13-s017_1.
[37] Synopsys® Formality® Version H-2013.03-SP5.
[38] H. Kalargaris, Y.-C Chen, and V. F. Pavlidis, "STA Compatible Backend Design Flow for TSV-based 3-D ICs," *Proceedings of the IEEE International Symposium on Quality Electronic Design,* pp. 186-191, March 2017.
[39] Synopsys® PrimeTime® Version H-2013.06-SP3-3.
[40] G. Katti *et al.*, "Electrical Modeling and Characterization of Through Silicon via for Three-Dimensional ICs," *IEEE Transactions on Electron Devices*, Vol. 57, No. 1, pp. 256-262, January 2010.
[41] *International Technology Roadmap for Semiconductors ITRS*, 2011.
[42] S. Davidson, "ITC99 Benchmark Circuits-Preliminary Results," *Proceedings of the IEEE International Test Conference,* pp. 1125-1125, September 1999.
[43] www.opencores.com
[44] S. Swaminathan *et al.*, "A Dynamically Reconfigurable Adaptive Viterbi Decoder," *Proceedings of ACM International Symposium on FPGAs,* pp. 227-236, February 2002.